fortiss

fortiss fields of research

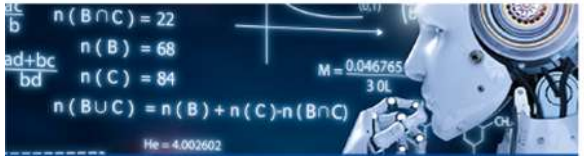
Chair of Software Engineering for Data-intensive Applications

Architectures and Services for Critical Infrastructures
Simple design and clear modeling for software simulation and integration

Automated Software Testing
Software Engineering for data intensive applications

Human-centered Engineering
Understand and explain descisions of intelligent systems from the user's point of view

Industrial Internet of Things
Enabling the next generation of IIoT applications and services



Machine Learning
Development of solutions involving data and knowledge

Model-based Systems Engineering
Solutions for flexible engineering of cyber-physical systems

Neuromorphic Computing
Artificial intelligence: the third generation of neural networks

Platform Engineering
Pervasive, robust and trustworthy platforms

Requirements Engineering
Efficiently deal with volatile and heterogenous requirements

Safety and Security
Guaranteeing secure systems in software and system development

Software Dependability
Rigorous validation and verification for dependable and safe software systems

Automated Software Testing

We like to generate tests, monitor them, and make them real!

Dr. Andrea Stocco
Head of AST @ fortiss
Prof. @ TUM

stocco@fortiss.org
andrea.stocco@tum.de

Stefano Carlo Lambertenghi
Generative AI Testing /
Reality Gap Assessment and Mitigation

lambertenghi@fortiss.org

Oliver Weißl
Algorithm Optimization
(co-supervised
w/ S. Kacianka)

weissl@fortiss.org

Lev Sorokin
Algorithm Optimization /
Cross-Simulation Testing

sorokin@fortiss.org

Xingcheng Chen
eXplainable Artificial
Intelligence (XAI) /
Post-Production Testing

xchen@fortiss.org

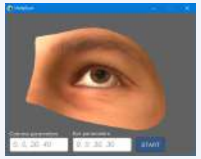
Davide Yi Xian Hu
Generative AI Testing
(visiting)

hu@fortiss.org

Automated Software Testing

Main domain areas

- **Traditional Software Systems**
  - Search-based Test Generation
  - Regression Testing (e.g., test prioritization, test minimization)
- **Web Applications**
  - End-to-End GUI Testing
  - Test Maintenance
  - Test Robustness
  - Automated Crawling
- **Deep Learning Systems**
  - Test Generation (search-based, model-based, generative AI)
  - Failure Prediction (black-box, white-box)
  - Debugging (explainable AI, uncertainty quantification)

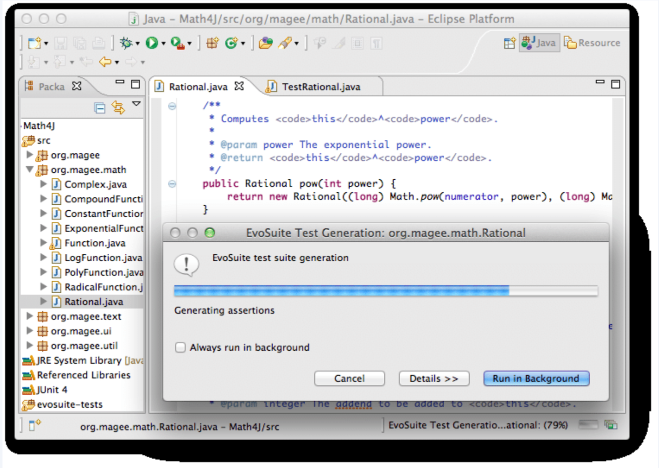


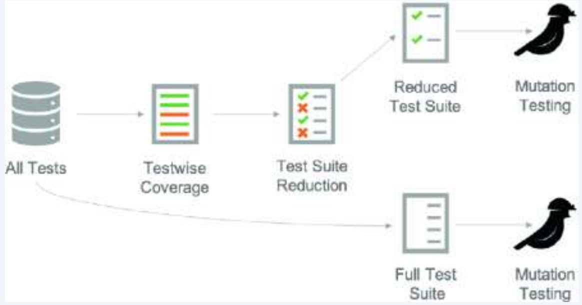
- **Cyber Physical Systems**
  - Deep learning or Reinforcement learning
  - Reality Gap Assessment and Mitigation
  - Automotive, UAVs, Elevators

Automated Software Testing

Main domain areas

- **Traditional Software Systems**
  - Search-based Test Generation
  - Regression Testing (e.g., test prioritization, test minimization)
- **Web Applications**
  - End-to-End GUI Testing
  - Test Maintenance
  - Test Robustness
  - Automated Crawling
- **Deep Learning Systems**
  - Test Generation (search-based, model-based, generative AI)
  - Failure Prediction (black-box, white-box)
  - Debugging (explainable AI, uncertainty quantification)
- **Cyber Physical Systems**
  - Deep learning or Reinforcement learning
  - Reality Gap Assessment and Mitigation
  - Automotive, UAVs, Elevators

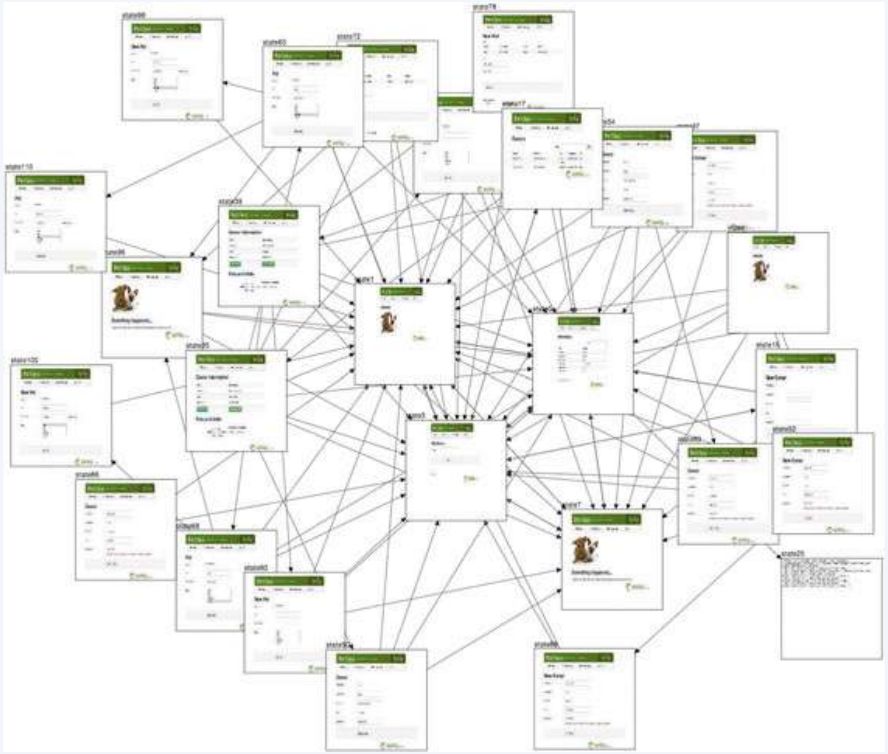




Automated Software Testing

Main domain areas

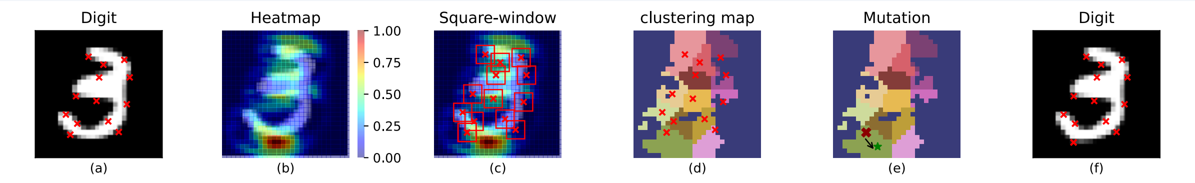
- **Traditional Software Systems**
  - Search-based Test Generation
  - Regression Testing (e.g., test prioritization, test minimization)
- **Web Applications**
  - End-to-End GUI Testing
  - Test Maintenance and Repair
  - Test Robustness
  - Automated Crawling
- **Deep Learning Systems**
  - Test Generation (search-based, model-based, generative AI)
  - Failure Prediction (black-box, white-box)
  - Debugging (explainable AI, uncertainty quantification)
- **Cyber Physical Systems**
  - Deep learning or Reinforcement learning
  - Reality Gap Assessment and Mitigation
  - Automotive, UAVs, Elevators

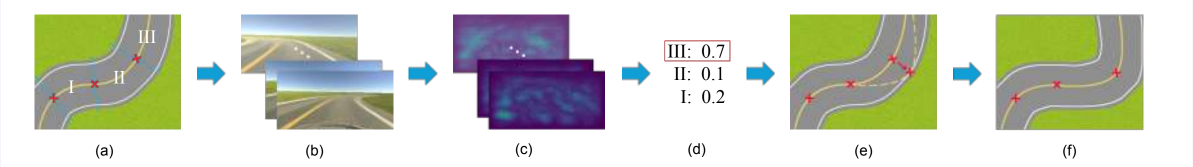


Automated Software Testing

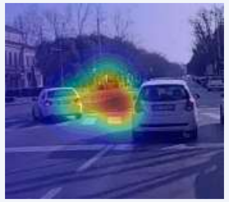
Main domain areas

- Traditional Software Systems
  - Search-based Test Generation
  - Regression Testing (e.g., test prioritization, test minimization)
- Web Applications
  - End-to-End GUI Testing
  - Test Maintenance
  - Test Robustness
  - Automated Crawling
- **Deep Learning Systems**
  - Test Generation (search-based, model-based, generative AI)
  - Failure Prediction (black-box, white-box)
  - Debugging (explainable AI, uncertainty quantification)
- Cyber Physical Systems
  - Deep learning or Reinforcement learning
  - Reality Gap Assessment and Mitigation
  - Automotive, UAVs, Elevators

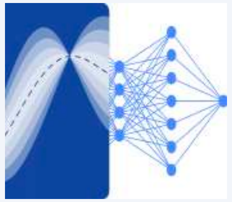




OOD detection



eXplainable AI



Uncertainty

Automated Software Testing

Main domain areas

- Traditional Software Systems
  - Search-based Test Generation
  - Regression Testing (e.g., test prioritization, test minimization)
- Web Applications
  - End-to-End GUI Testing
  - Test Maintenance
  - Test Robustness
  - Automated Crawling
- Deep Learning Systems
  - Test Generation (search-based, model-based, generative AI)
  - Failure Prediction (black-box, white-box)
  - Debugging (explainable AI, uncertainty quantification)
- **Cyber Physical Systems**
  - Deep learning or Reinforcement learning
  - Reality Gap Assessment and Mitigation
  - Automotive, UAVs, Elevators

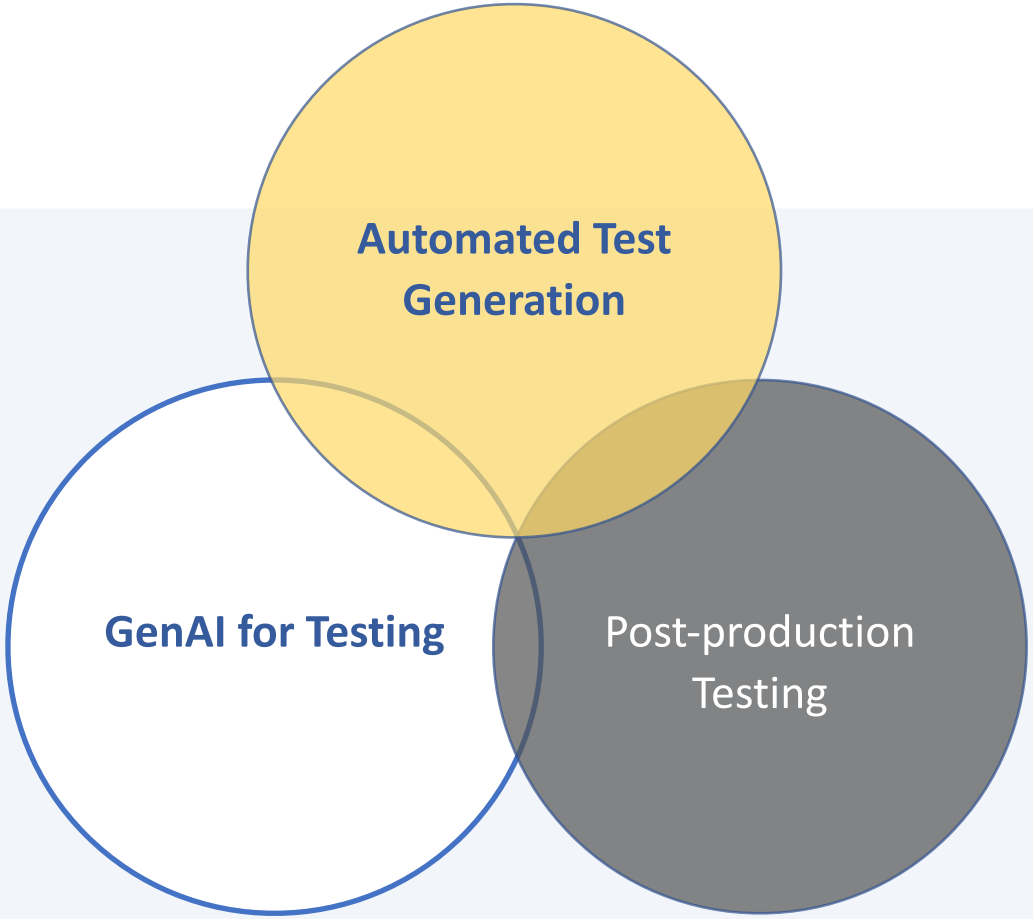
Automated Software Testing

Main research topics

- **Automated Test Generation**
  How can we automatically generated complex scenario-based tests efficiently and effectively?

- **GenAI for Testing**
  How can we leverage generative adversarial techniques, uncertainty quantification and conformal predictions, explainable AI for testing CPS

- **Post-production Testing**
  How to ensure a high dependability of deep neural network driven-cyber-physical systems (CPS) in production?



Automated Software Testing

Search-based Evolutionary Algorithms

ADS & ADAS

- End-to end DNNs (level 2)
- Full AD stacks

Critical Test Scenarios Identification

High fidelity sim-based Validation

CARLA UDACITY BeamNG

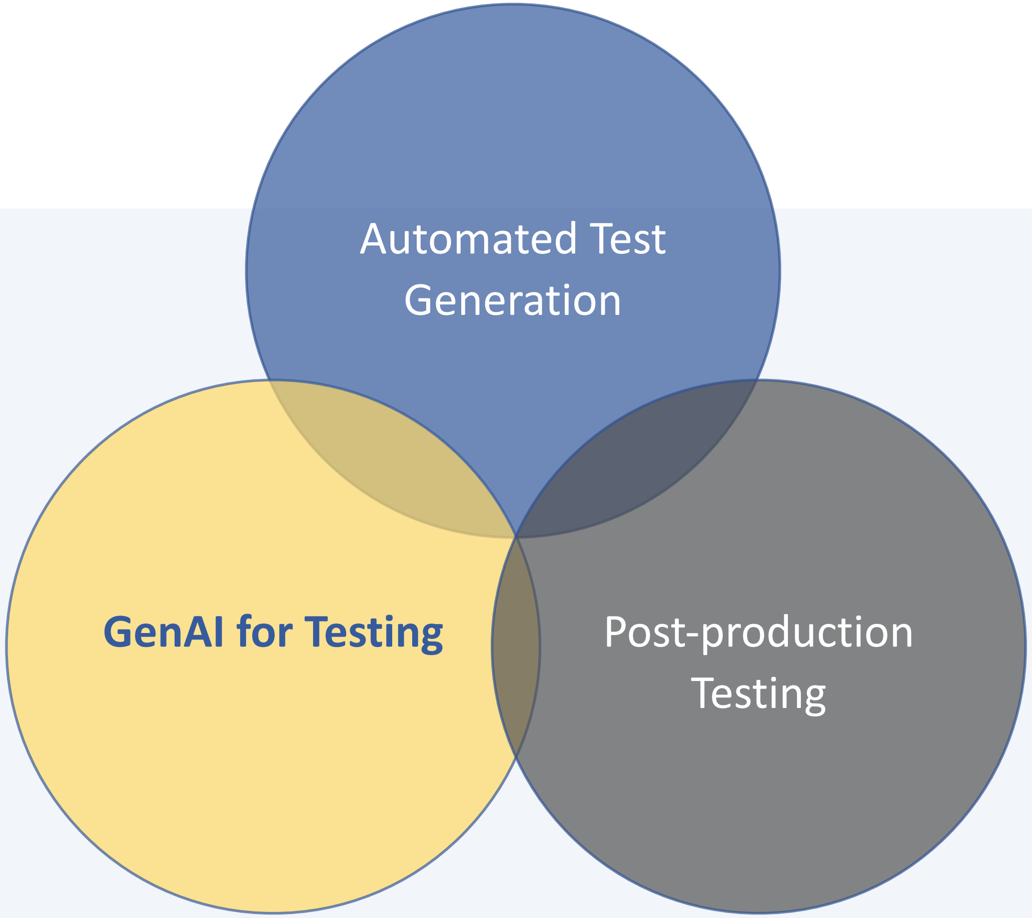
https://git.fortiss.org/opensbt

Sorokin et al. OpenSBT: A Modular Framework for Search-based Testing of Automated Driving Systems. In Proceedings of ICSE Workshops 2024.

Automated Software Testing

Main research topics

- **Automated Test Generation**
  How can we automatically generated complex scenario-based tests efficiently and effectively?
- **GenAI for Testing**
  How can we leverage generative adversarial techniques, uncertainty quantification and conformal predictions, explainable AI for testing CPS
- **Post-production Testing**
  How to ensure a high dependability of deep neural network driven-cyber-physical systems (CPS) in production?



Reality Gap

Our mitigation and evaluation techniques

Generic simulators

CARLA UDACITY

Generic simulators + Vehicle Digital twin

Improved vehicle dynamics and sensors

Generic simulators + Real software stack

apollo

Improved vehicle behaviour

Generative models for perception inputs (offline)

CYCLEGAN/ Pix2Pix

Improved sensor fidelity

Small scale vehicle

Improved vehicle behaviour and physics

Small scale vehicle + Mixed reality

Improved behaviour and scenarios

Scenario generation from real data

Improved scenarios

Generative models for simulation rendering

Improved sensor fidelity

Reality Gap

Our mitigation and evaluation techniques

Generic simulators

CARLA UDACITY

Improved ...

Generic simulators + Vehicle Digital twin

Improved vehicle dynamics and sensors

Generic simulators + Real software stack

apollo

Improved vehicle behaviour

Generative models for perception inputs (offline)

CYCLEGAN/ Pix2Pix

Improved sensor fidelity

Small scale vehicle

Improved vehicle behaviour and physics

Small scale vehicle + Mixed reality

Improved behaviour and scenarios

Scenario generation from real data

Improved scenarios

Generative models for simulation rendering

Improved sensor fidelity

Perception Reality Gap

Difference between simulated and real input images

Simulation

Real-world

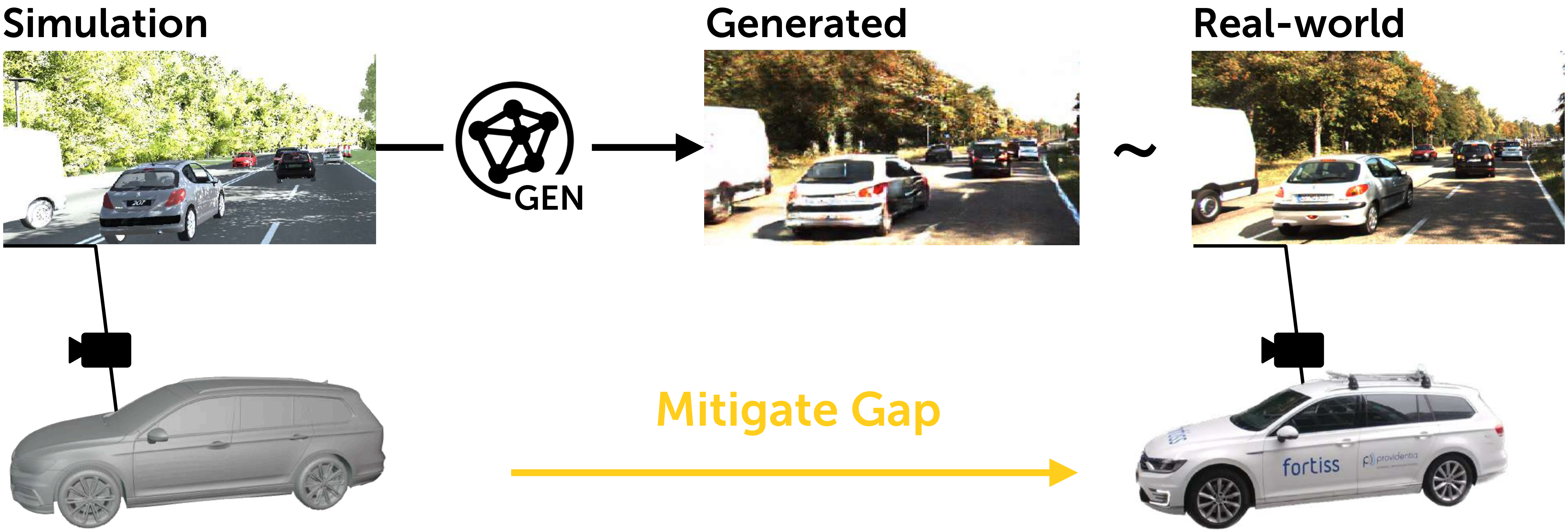
Perception Gap

ADS **Simulated Behaviour** ≠ ADS **Real Behaviour**

Gaidon, A et al. 2016
Geiger, A et al. 2013

Generative Image-to-Image Translation

Generative models for perception reality gap mitigation




REWE

Generative Image-to-Image Translation shortcomings

Generated

Real-world

Methodology

Image Quality Metrics

Distribution Level Metrics

IS	Inception-score
FID	Fréchet Inception Distance
KID	Kernel Inception Distance

Single Image Metrics

SSIM	Structural Similarity Index
PSNR	Peak signal-to-noise ratio
MSE	Mean Squared Error
CS	Cosine Similarity
TSI	Texture Similarity Index
WD	Wasserstein Score
KL	KL Divergence
HistI	Histogram Intersection
CPL	Classifier Perceptual Loss
SSS	Semantic Segmentation Score

Borji, A. et al. 2018
Pang, Y. et al. 2022

Empirical evaluation

Image Metrics VS Behaviour Metrics

Correlation

How do existing Image-to-image evaluation metrics correlate with the associated ADS behaviour?

Takeaways

GEN

REAL

Relative Behaviour Metrics ADS

Image Metrics GEN

1. Image-to-image GenAI tools effectively tackle domain adaptation in ADS
2. Current GenAI metrics don't align well with the software behavior that relies on their output
3. We need more domain-informed, semantic-aware metrics

Lambertenghi and Stocco.
Assessing Quality Metrics for Neural Reality Gap Input Mitigation in Autonomous Driving Testing
17th IEEE International Conference on Software Testing, Verification and Validation (ICST) 2024

Reality Gap

Our mitigation and evaluation techniques

Generic simulators

CARLA UDACITY

Generic simulators + Vehicle Digital twin

Improved vehicle dynamics and sensors

Generic simulators + Real software stack

apollo
THE AUTOWARE FOUNDATION

Improved vehicle behaviour

Generative models for perception inputs (offline)

CYCLEGAN/ Pix2Pix

Improved sensor fidelity

Small scale vehicle

Improved vehicle behaviour and physics

Small scale vehicle + Mixed reality

Improved behaviour and scenarios

Scenario generation from real data

Improved scenarios

Generative models for simulation rendering

\+

Improved sensor fidelity

Gap mitigation

Simulation testing

Simulated

Behaviour
Scenarios
Sensors

Small-scale
Real-world testing

Real

Behaviour
Sensors

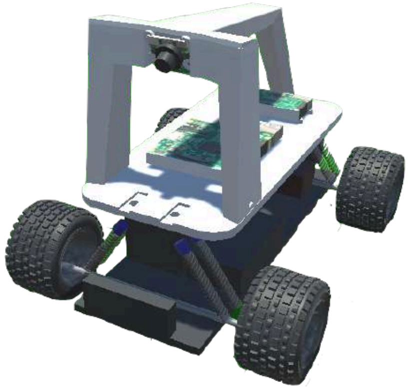
Full-scale
Real-world testing

Real

Behaviour
Scenarios
Sensors

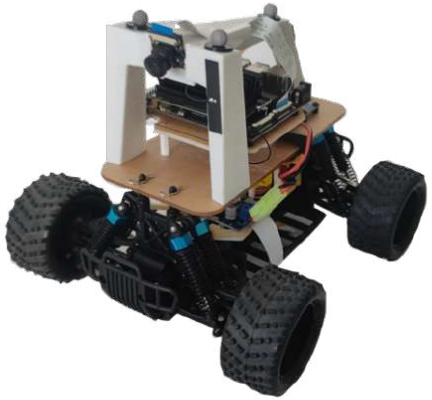
Gap mitigation

Simulation testing



+

Real-world testing



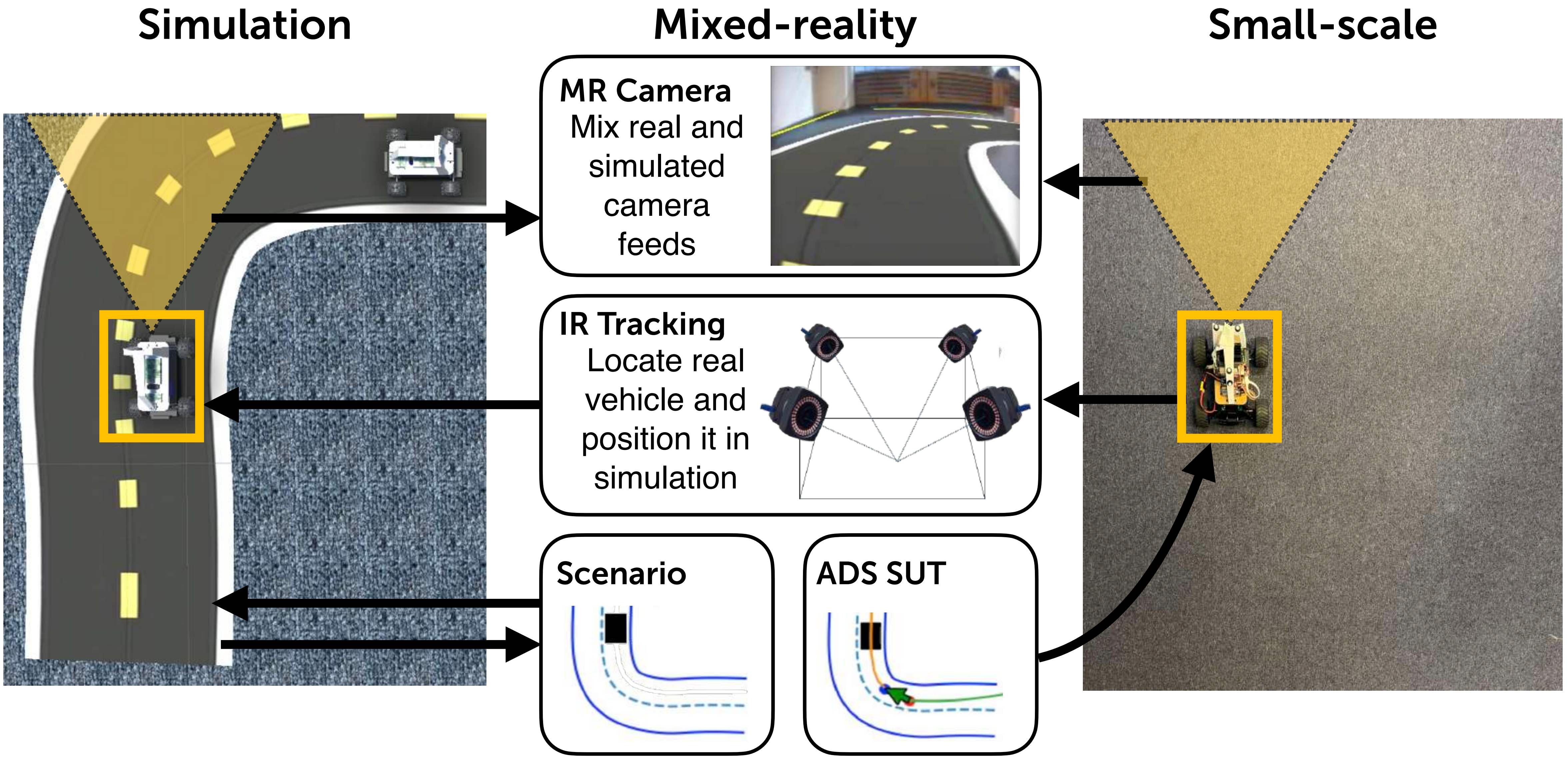
=

Mixed-reality

⊕ More realistic than sim

⊕ More versatile than full-scale

Mixed-reality framework



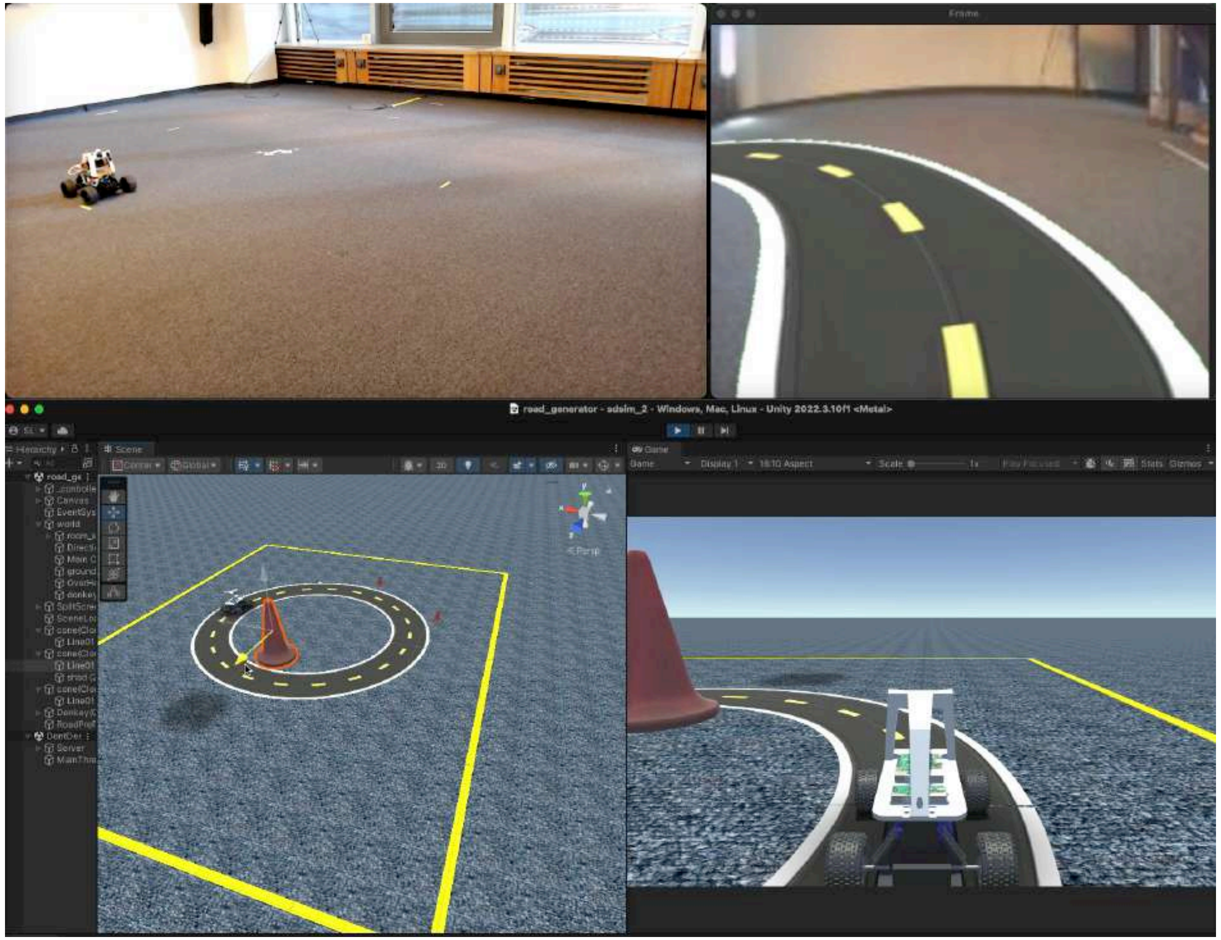
Mixed-reality testing framework

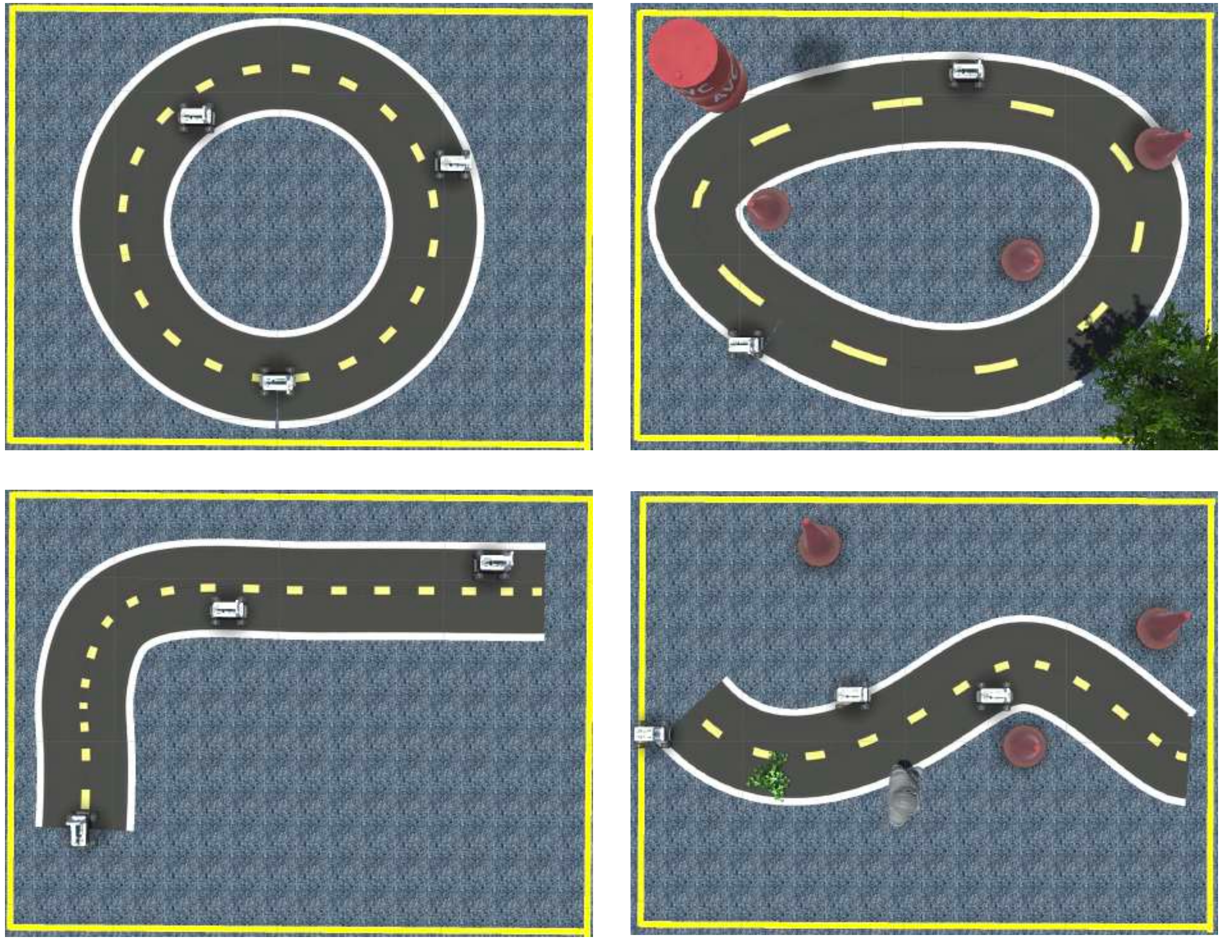
unity

- Test complex scenarios
- Track obstacles in real-time
- End-to-end ADS testing
- Full-stack ADS testing

Mixed-reality framework

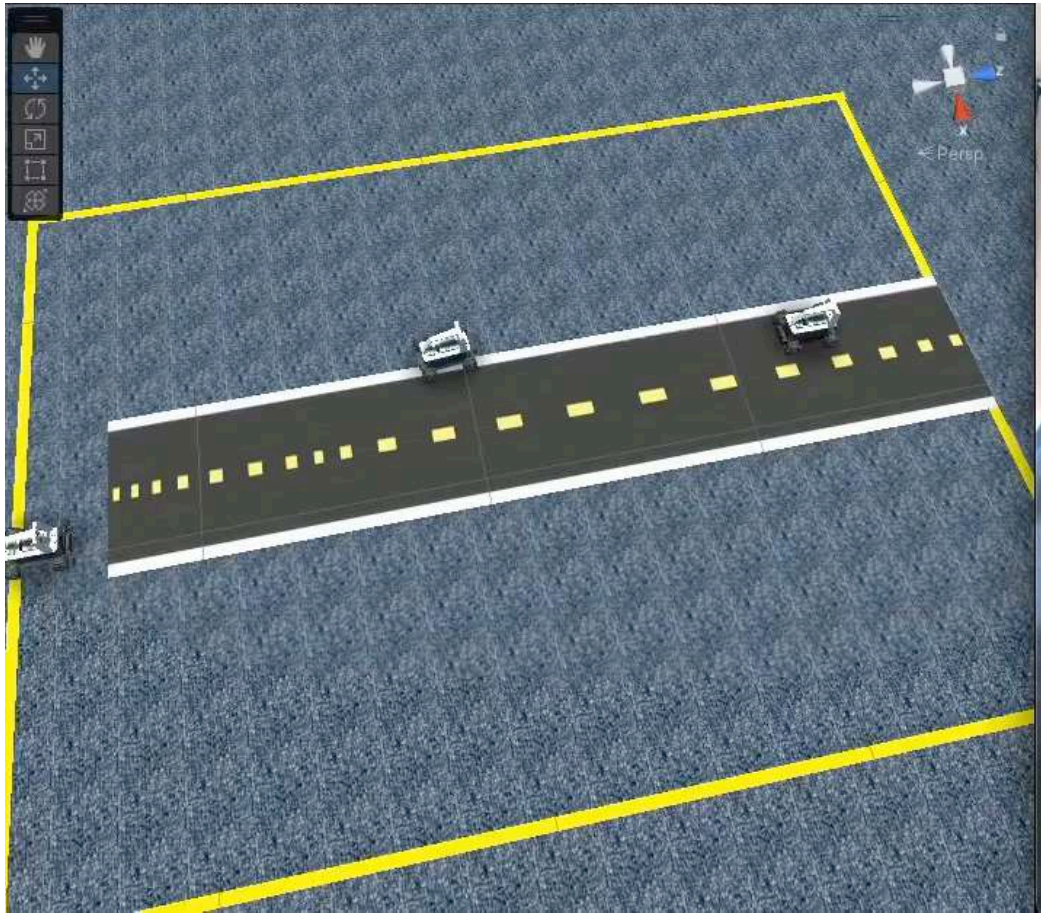
- Test complex scenarios





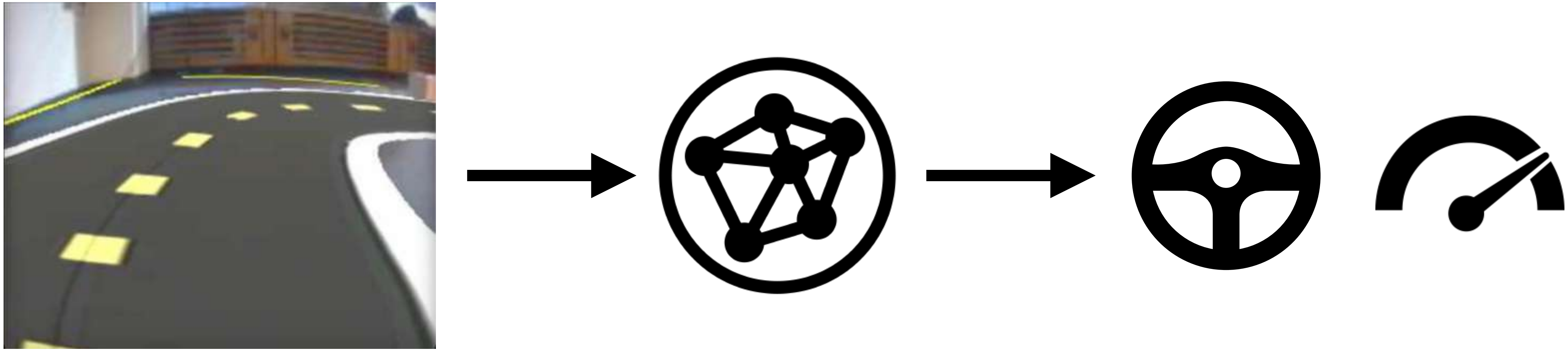
Mixed-reality framework

- Track obstacles in real-time



Mixed-reality framework

End-to-end ADS execution



Mixed feed

Lane-keeping ADAS (DNN)

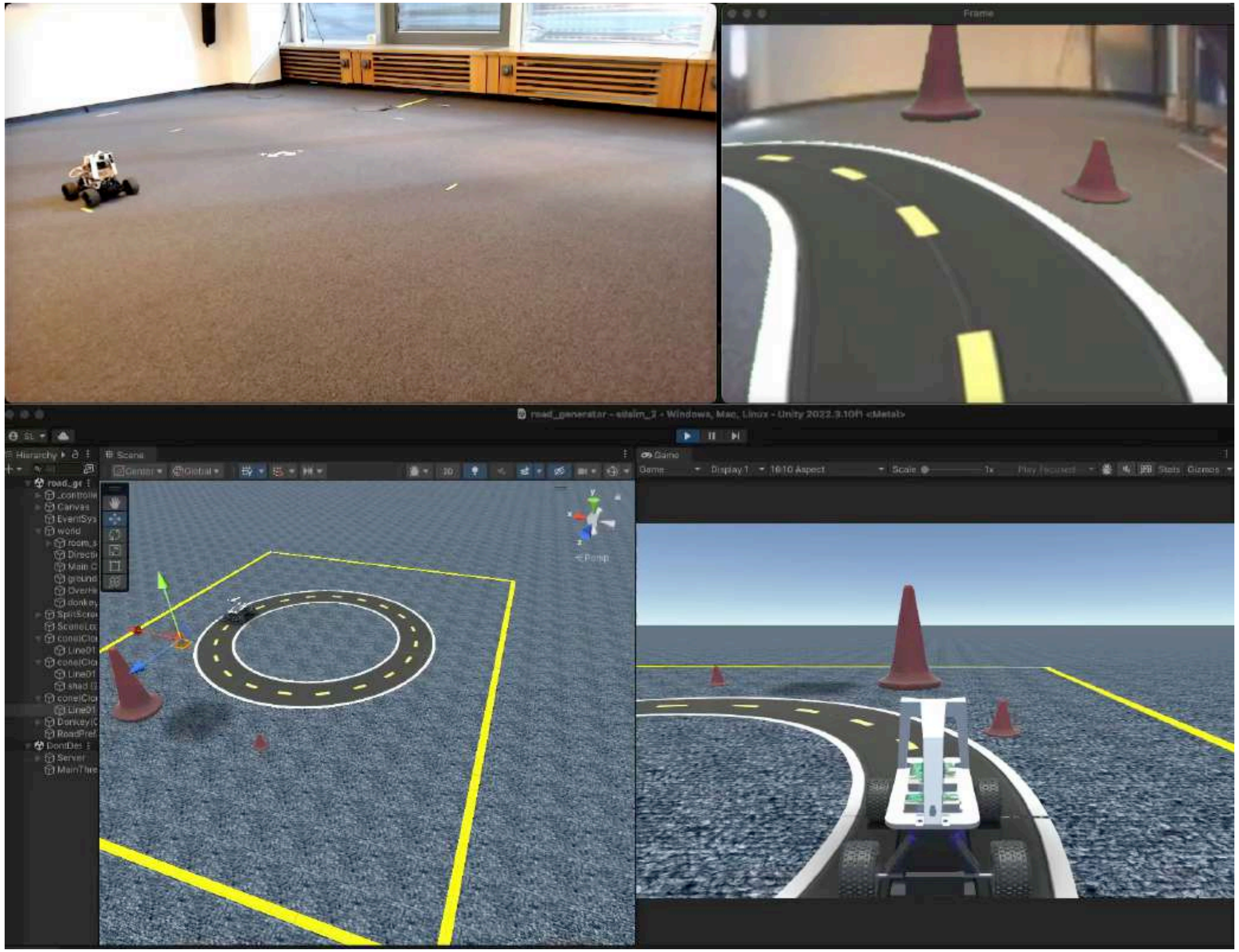
Outputs

Mixed-reality framework

End-to-end ADS execution

Real world

Mixed-reality

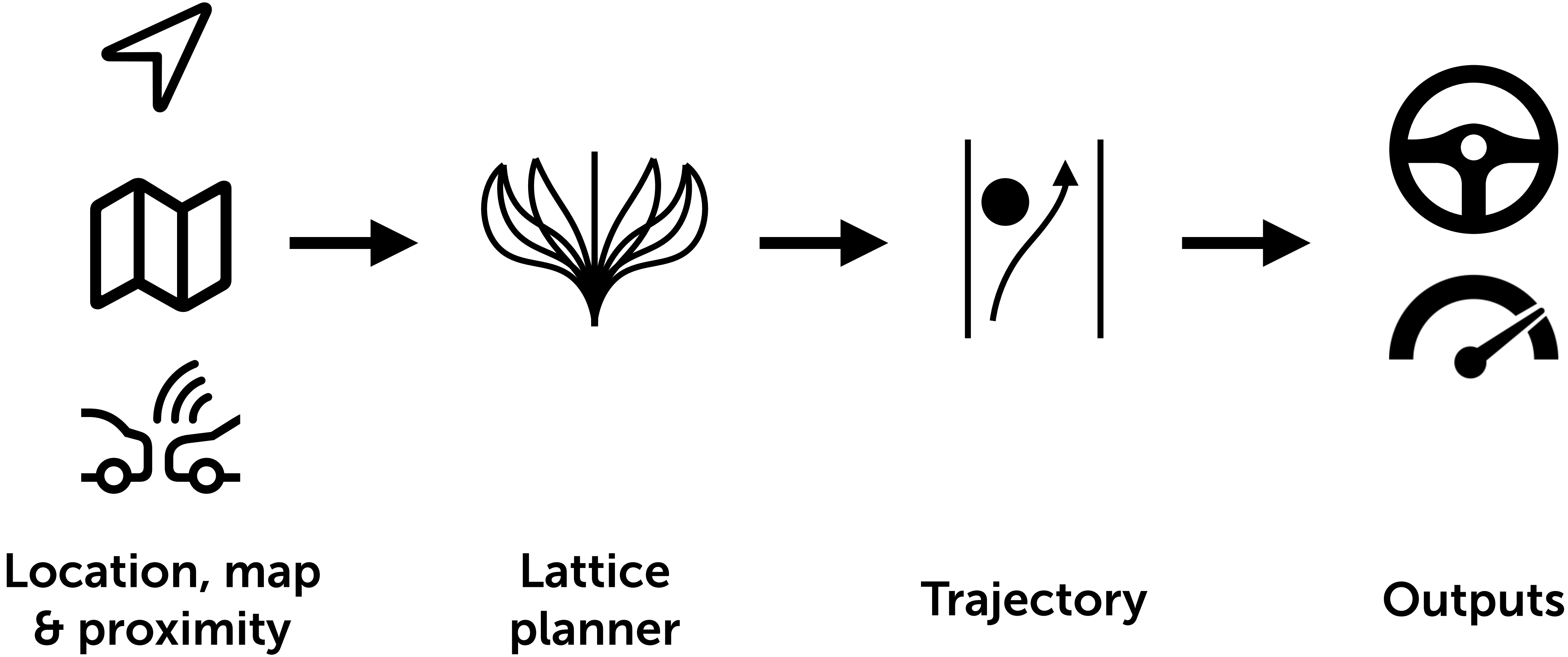


Simulation

Simulation
(FOV)

Mixed-reality framework

Path-planning ADS execution



Mixed-reality framework

unity

+ Real software stack

Location, map, camera & LIDAR

apollo

THE AUTOWARE FOUNDATION

Outputs

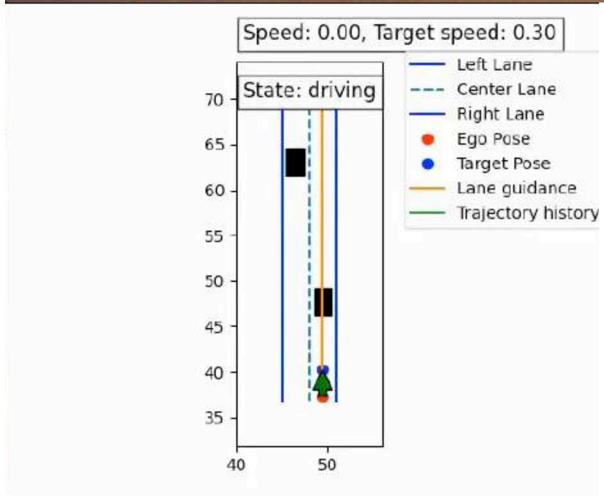
Mixed-reality framework

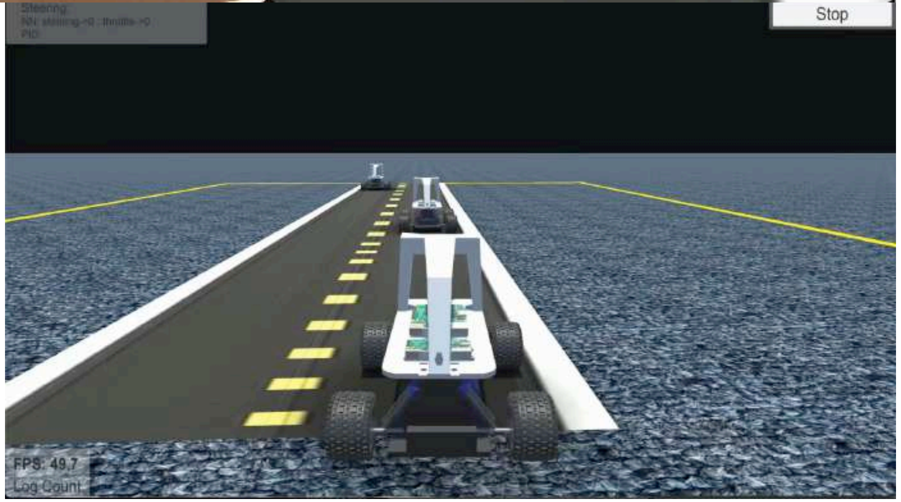
Path-planning ADS execution

Real world

Mixed-reality

Path planning





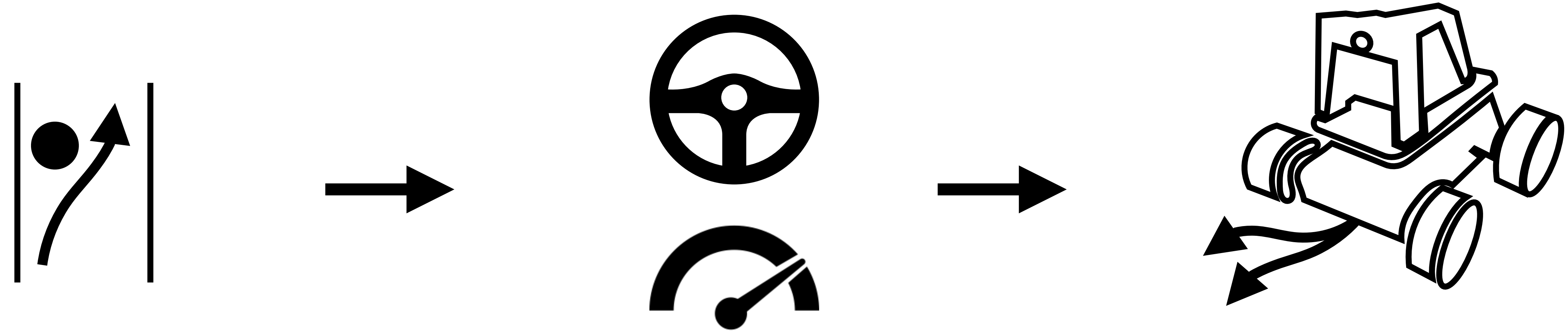
Simulation
(FOV)

Mixed-reality framework

- + Autoware Planning

Demonstration

A small sim2real gap demonstration



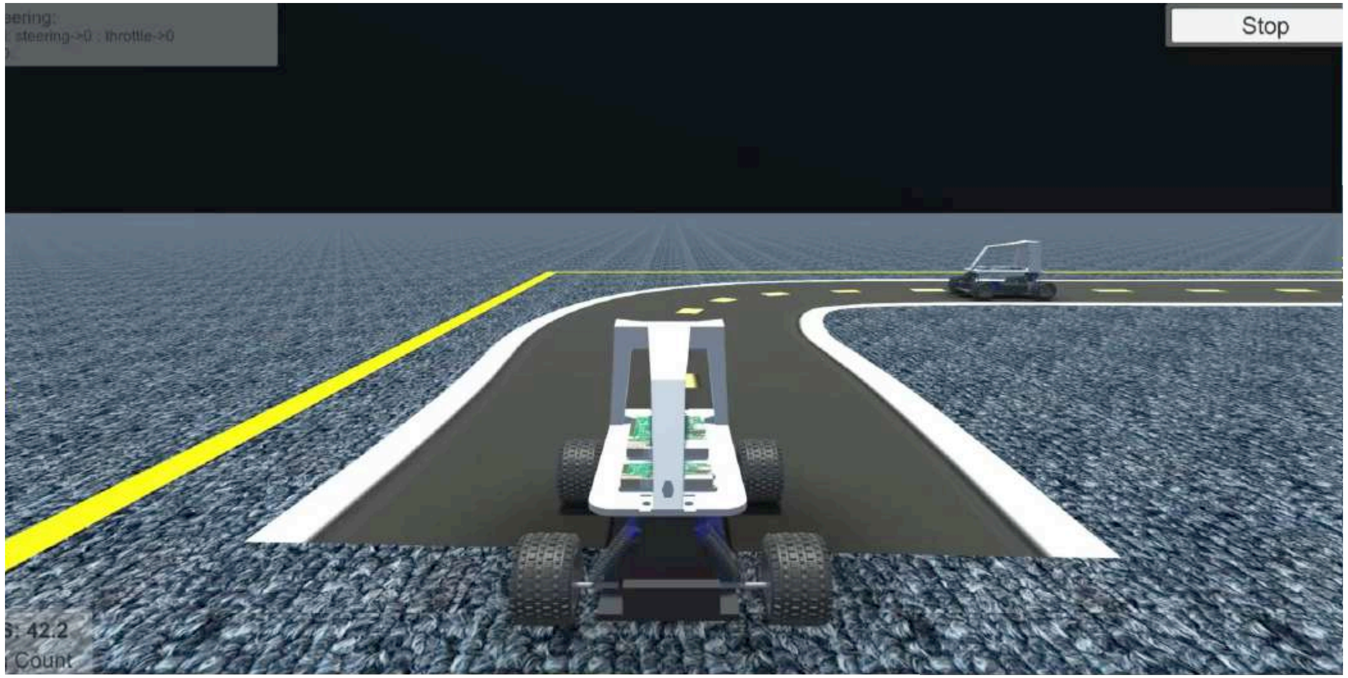
Trajectory

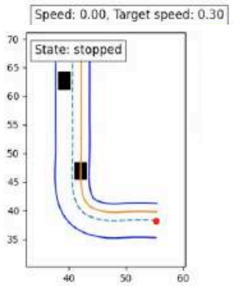
Outputs

Sim actuation
≠
Real actuation

Demonstration

Simulation only



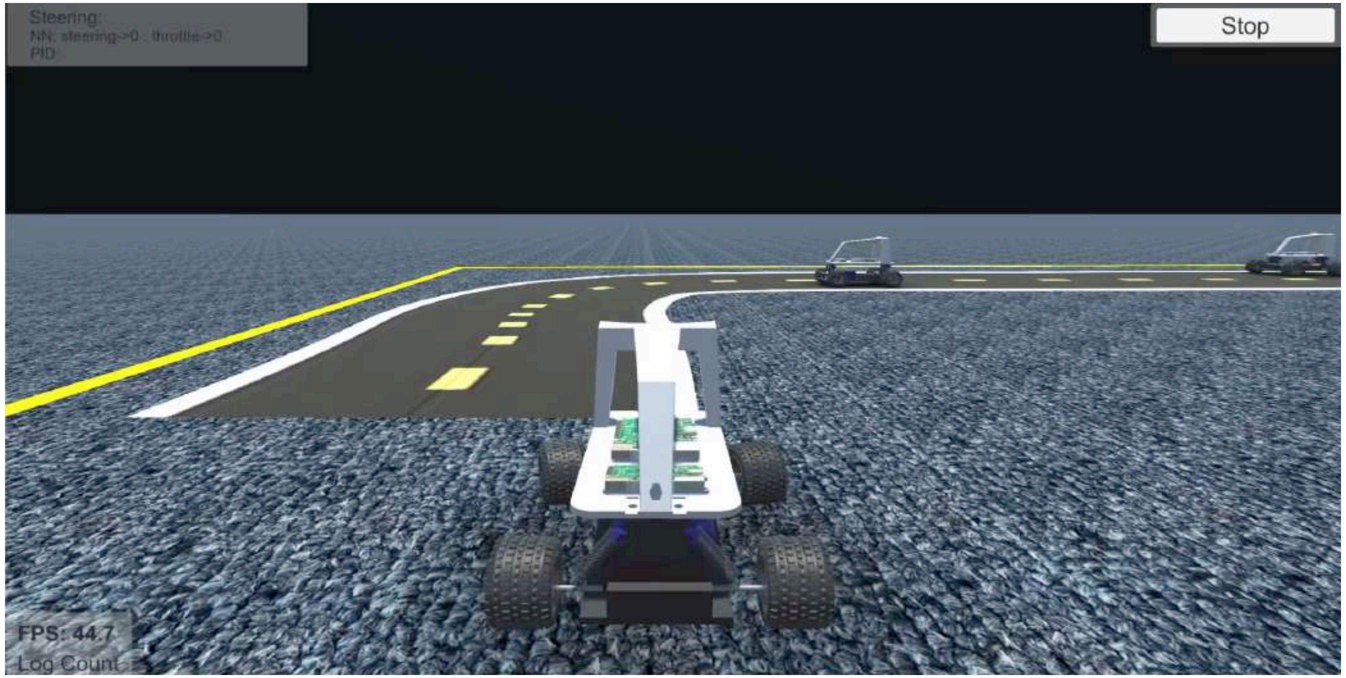


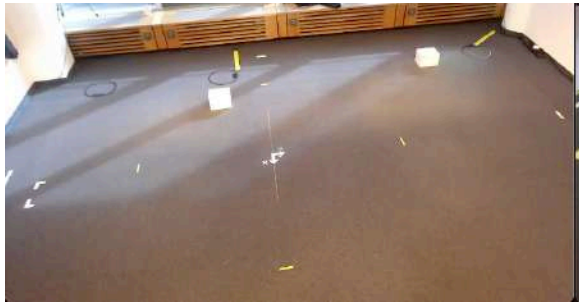
Safe distance: 0.5m

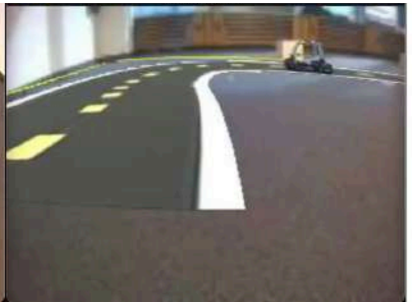
Target speed: 0.3m/s

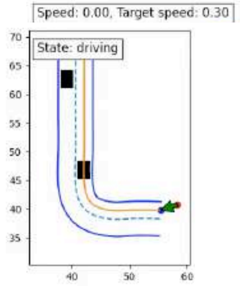
Demonstration

Mixed-reality execution







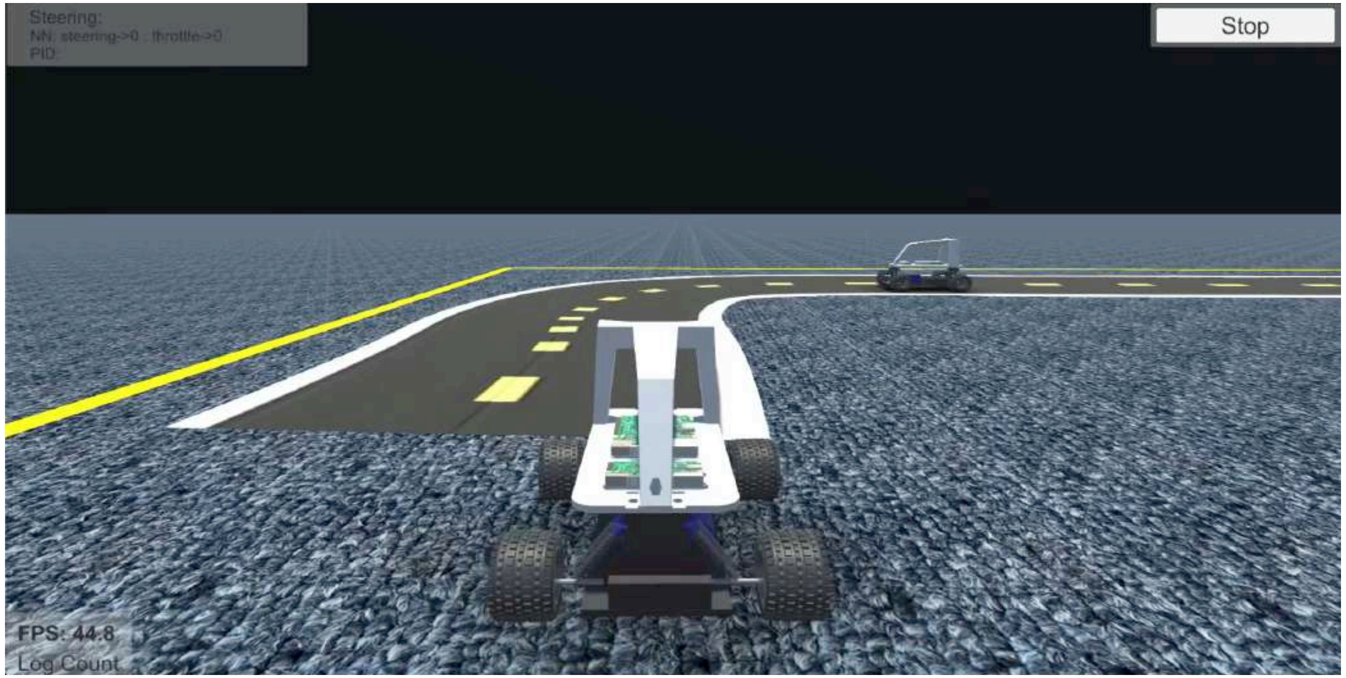


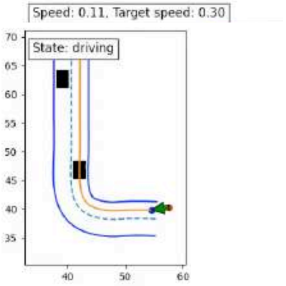
Safe distance: 0.5m

Target speed: 0.3m/s

Demonstration

Rectified mixed-reality execution





Safe distance: 0.7m

Target speed: 0.3m/s

Reality Gap

Our mitigation and evaluation techniques

Generic simulators

CARLA UDACITY

Generic simulators + Vehicle Digital twin

Improved vehicle dynamics and sensors

Generic simulators + Real software stack

apollo

Improved vehicle behaviour

Generative models for perception inputs (offline)

CYCLEGAN/ Pix2Pix

Improved sensor fidelity

Small scale vehicle

Improved vehicle behaviour and physics

Small scale vehicle + Mixed reality

Improved behaviour and scenarios

Scenario generation from real data

Improved scenarios

Generative models for simulation rendering

+

Improved sensor fidelity

ADAS requires extensive coverage of the ODD

From regulations to implementation

Existing Standards and Regulations

- ISO/PAS 21448 Safety of the Intended Function (SOTIF)
- UN Regulation No 157 (2021/389)
- ISO 34505 “Scenery Elements (Section 9)” and “Environmental Conditions (Section 10)”

Operational Design Domain (ODD)

- roadway types
- geographic area
- environmental conditions (weather as well as day/night time)

Diffusion Models

Used for Training Set Augmentation

Augmentation: **Lightning Strikes**


Input Image


Instruction-edited


Inpainting


Inpainting with Refining

Augmentation: **Autumn Season**


Input Image

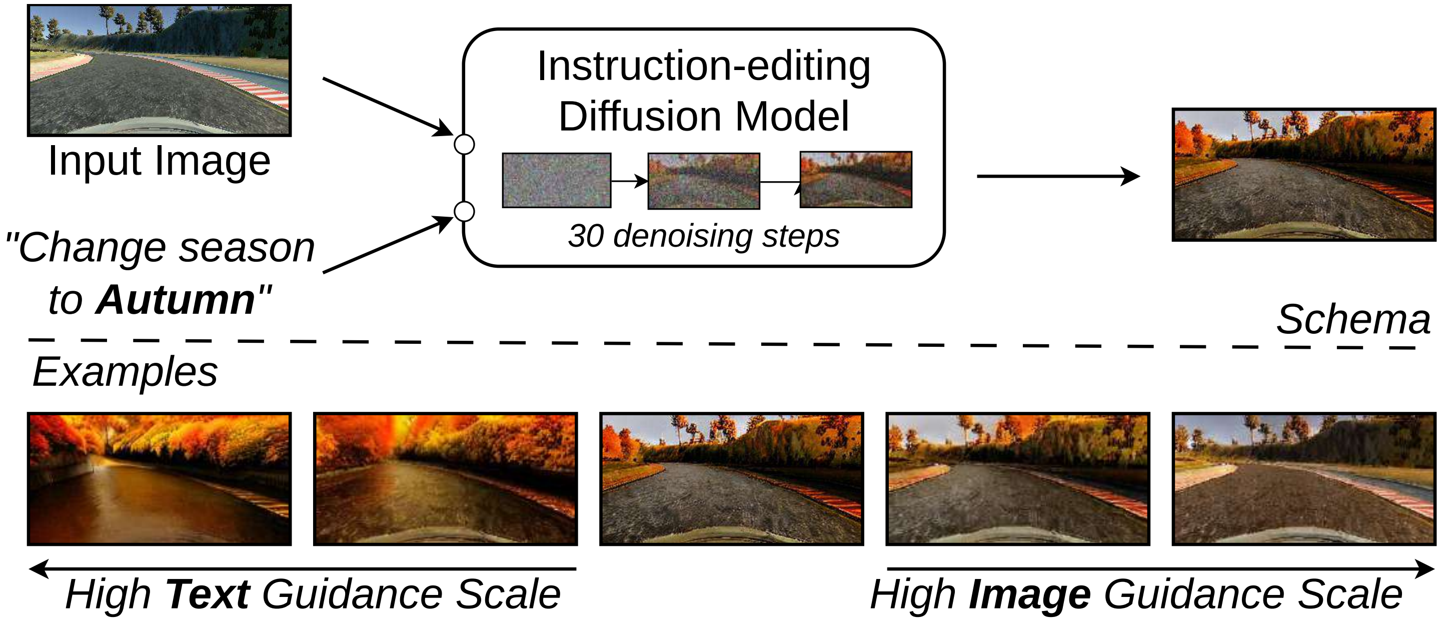

Instruction-edited


Inpainting


Inpainting with Refining

Instruction-editing

Prompt: Textual



Enhancing ADS Testing with Driving Simulators and Generative AI

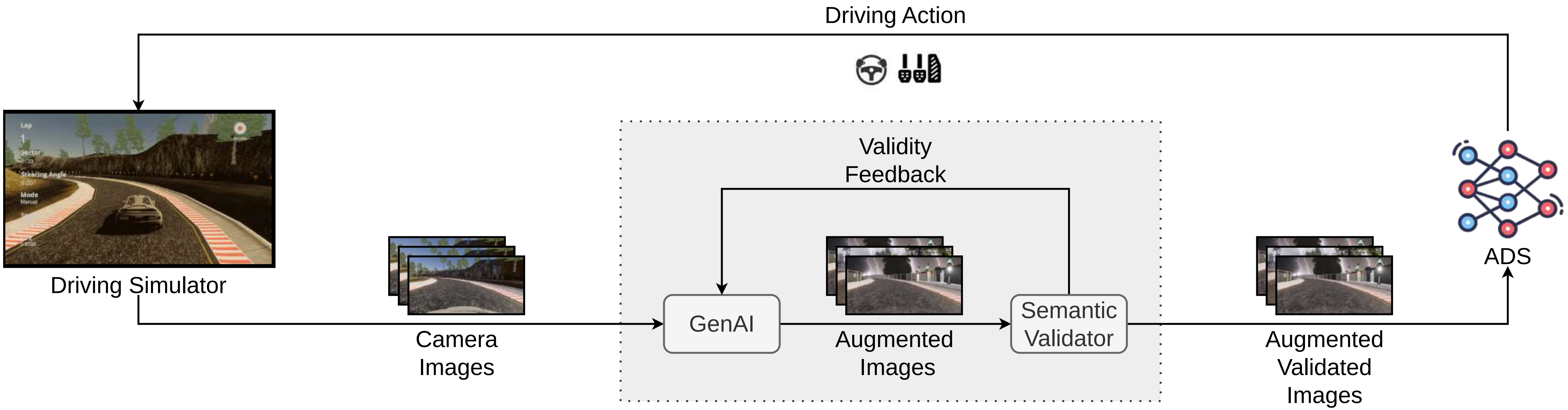
Simulators with Generative AI (naïve integration)

InstructPix2Pix

- ✓ Diversity
- ✗ Temporal Consistency
- ✗ Efficiency +11-20X overhead

Enhancing ADS Testing with Driving Simulators and Generative AI

Simulators with Generative AI



Enhancing ADS Testing with Driving Simulators and Generative AI

Simulators with Generative AI (knowledge distillation)

InstructPix2Pix with Knowledge Distillation

- Diversity
- Temporal Consistency
- Efficiency +0.02X overhead

Baresi, Hu, Stocco, Tonella.
Efficient Domain Augmentation for Autonomous Driving Testing Using Diffusion Models.
47th IEEE/ACM International Conference on Software Engineering (ICSE) 2025


Egypt
Brazil
Australia
Argentina
Canada
Germany
Evening
Sunset
Dawn
England
Beijing
Morning
Twilight
Autumn
Berlin
Night
Sunrise
Dusk
Dust
Dubai
Forest
Blizzard
Snow
Foggy

Takeaways

ORIG
ODD

NEW
ODD

ADS Behaviour Metrics

GEN Diffusion Models

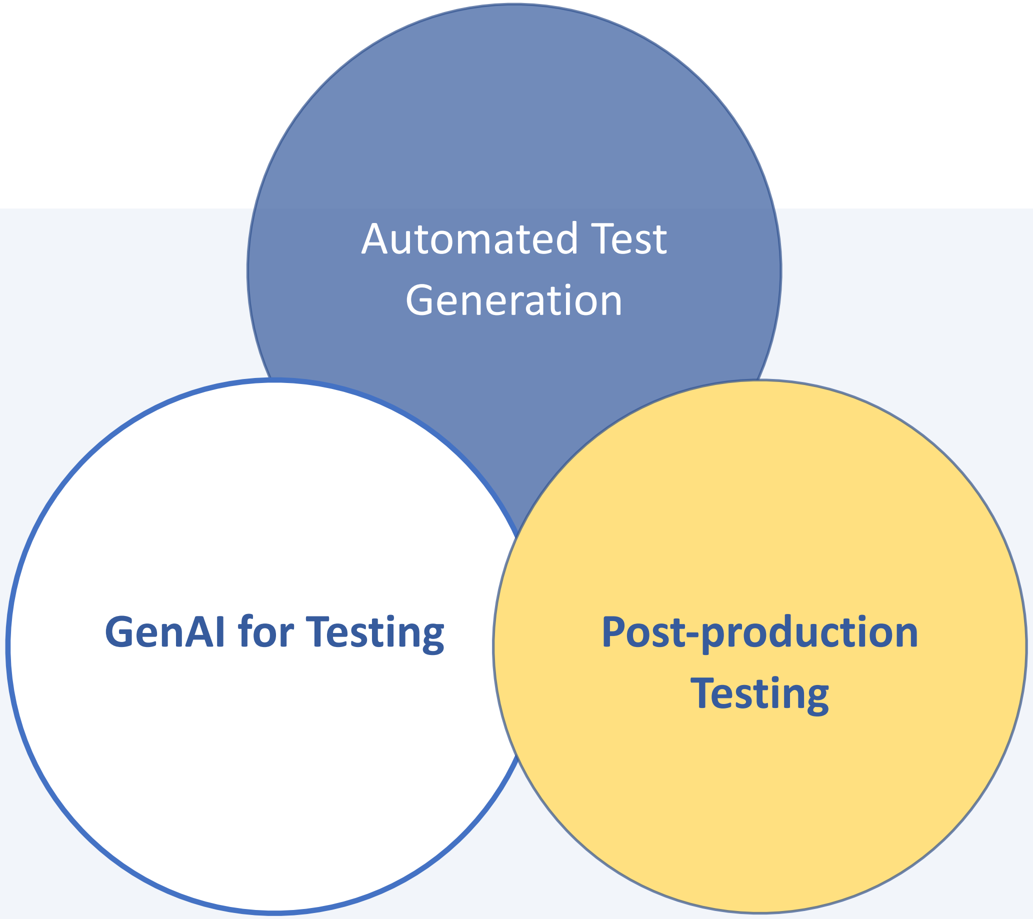
1. **Diffusion models effectively tackle domain generation for ADS testing**
2. **They complement simulator testing, uncovering failures in areas previously considered error-free**
3. **Knowledge distillation is key to achieving high simulation efficiency**

Baresi, Hu, Stocco, Tonella.
Efficient Domain Augmentation for Autonomous Driving Testing Using Diffusion Models.
47th IEEE/ACM International Conference on Software Engineering (ICSE) 2025

Automated Software Testing

Main research topics

- **Automated Test Generation**
  How can we automatically generated complex scenario-based tests efficiently and effectively?
- **GenAI for Testing**
  How can we leverage generative adversarial techniques, uncertainty quantification and conformal predictions, explainable AI for testing CPS
- **Post-production Testing**
  How to ensure a high dependability of deep neural network driven-cyber-physical systems (CPS) in production?



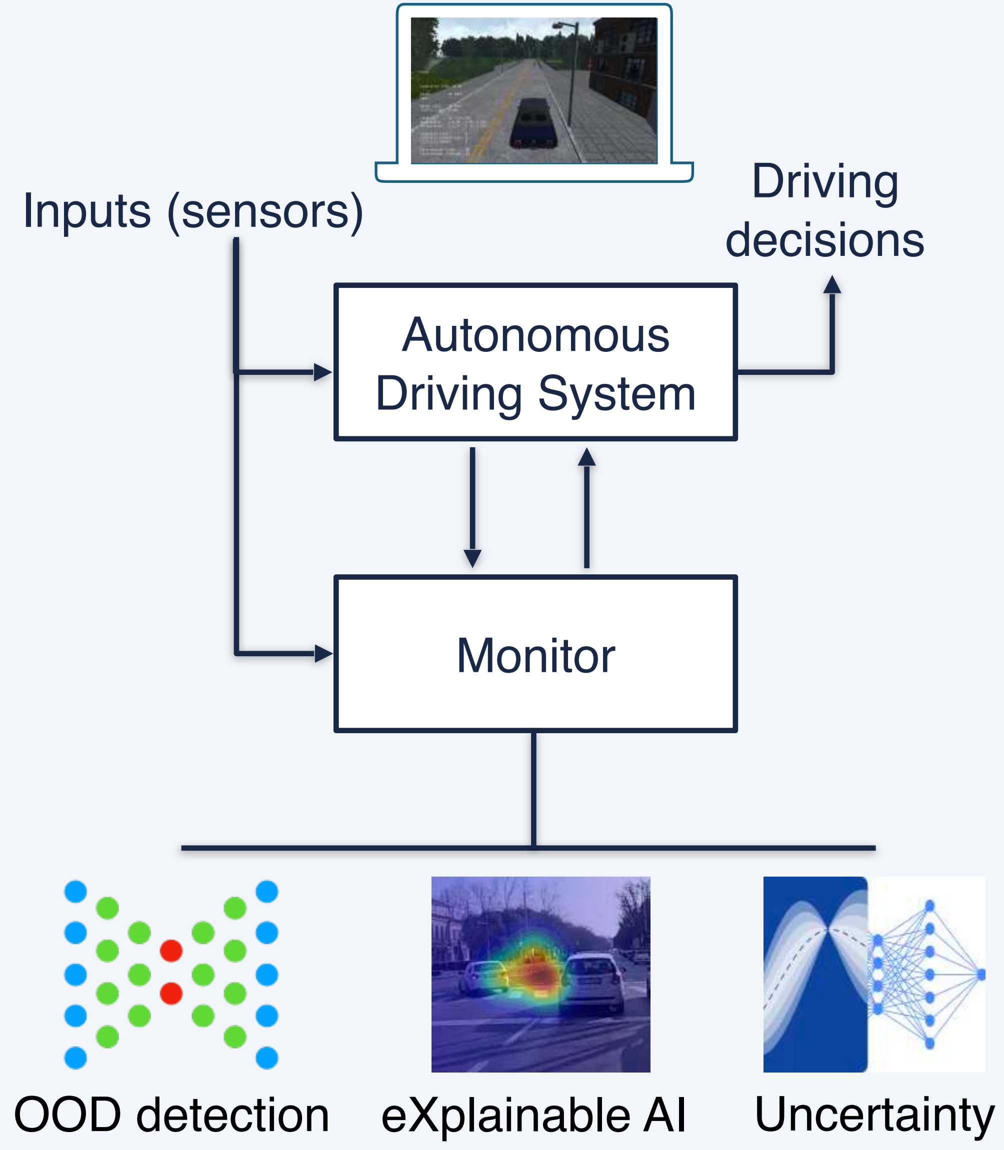
Post-production Testing

Real-time monitoring of deployed ADS

ADS & ADAS
- **End-to-end DNNs (level 2)**
- **Full AD stacks**

Real-time Monitoring
Observe System Under Test at runtime

Trustworthy ADS
Alert the driver if system is not to be trusted or activate fallback procedures



Stocco, Weiss, Calzana, Tonella.
Misbehaviour Prediction for Autonomous Driving Systems.
In Proceedings of ICSE 2020.

Vielen Dank!